DOCUMENT RESUME

ED 275 145 FL 016 043

AUTHOR Brown, James Dean
TITLE A Cloze Is a Cloze Is a Cloze?
PUB DATE Mar 83
NOTE 12p.; In: Handscombe, Jean, Ed.; And Others. On TESOL '83. The Question of Control. Selected Papers from the Annual Convention of Teachers of English to Speakers of Other Languages (17th, Toronto, Canada, March 15-20, 1983); see FL 015 035.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cloze Procedure; College Students; Comparative Analysis; *English (Second Language); Higher Education; *Item Sampling; *Language Proficiency; *Language Tests; Second Language Instruction; Standardized Tests; *Test Construction; Test Reliability; Test Validity

ABSTRACT

This study attempted to determine the effectiveness of cloze procedures as norm-referenced instruments by comparing the differential responses of four groups of college students of English as a second language on two identical cloze passages. The responses were scored using both exact-answer and acceptable-word methods. The results indicate that the effectiveness, measured as reliability and validity, appears to be strongly related to how well a given cloze passage fits a given student sample. This suggests, in turn, that pretesting any cloze passage is necessary so that an appropriate passage can be selected and modified or tailored to fit a certain group of students. Taking some or all of these steps should help produce a more reliable and valid norm-referenced instrument on whose scores responsible decisions about students can be based. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275145

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Frank McNeil

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Cloze is a Cloze is a Cloze?

James Dean Brown

INTRODUCTION

In 1953, Taylor first discussed a procedure, whereby some of the words in a written text were replaced by blanks and students were required to fill them in. He called this procedure "cloze" from the Gestalt psychology notion of closure, or human ability to fill gaps. Since then, there has been an explosion of research on cloze when applied to native English speakers and, more recently, on its utility among nonnative students of English (for overviews of this research, see Oller 1975; Oller 1979: 340-80).

For native speakers, cloze was originally designed by Taylor as a measure of the readability of texts. A great deal of work followed on this aspect of cloze (Oller 1979: 348-54). As an offshoot of this readability research, a number of studies have also been produced on cloze procedure as a measure of native-speaker reading comprehension ability (Brown 1978: 12-14). Criterion-related validity coefficients were calculated between cloze and various standardized reading tests in these studies. They ranged from .25 to .95. The squared values for these coefficients, .06 to .90, indicate the percent of shared, or overlapping, variance between cloze and the reading test in each study. It is safe to conclude from these results that cloze has been shown to be both a very weak (6 percent) and highly valid (90 percent) test of reading comprehension for native speakers—and almost everything in between as well.

For nonnatives, much of the work has been done on the value of cloze as a test of overall second language proficiency. Often studies focus on one or both of the key characteristics of a test: reliability and validity. For instance, studies have shown that cloze can be fairly reliable, that is, it produces consistent results. Such studies have indicated reliability indices ranging from .53 to .96 for various cloze passages (Darnell 1970; Oller 1972b; Pike 1973; Jonz 1976; Alderson 1979; Mullen 1979; Brown 1980; Hinofotis 1980; Brown 1983). Reliability coefficients can be interpreted as the percentage of reliable (consistent) variance in a test. Thus, cloze passages have been shown to have weak reliability (53 percent reliable variance) as well as high reliability (96 percent reliable variance)—and almost everything in between as well.

FL 016 043

J.D. Brown, having just completed two years with the UCLA/China Exchange Program at Zhongshan University in the PRC, is presently teaching at Florida State University.

The validity of cloze in second language situations has also been investigated (Conrad 1970; Darnell 1970; Oller and Inal 1971; Oller 1972 a and b; Irvine et al 1974; Stubbs and Tucker 1974; Alderson 1979 and 1980; Brown 1980; Hinofotis 1980; Mullen 1979). Validity is defined as the degree to which a test measures what it claims to be measuring—in this case, overall second language proficiency. Generally, this has been demonstrated (as criterion-related validity) by showing the strength of association between scores on a cloze test and those on a standardized language placement or proficiency examination. Coefficients of .43 to .91 have been reported in these studies. And again, the squared values of these coefficients, .19 to .83, indicate the percent of shared, or overlapping, variance between a given cloze test and the criterion measure. Hence, cloze has been shown to be a weak (19 percent) measure of overall language proficiency, as well as a fairly strong one (83 percent)—and almost everything in between as well.

It appears, then, that the results of studies on the reliability and validity of cloze procedure have varied greatly over the years. And in all fairness, it should be pointed out that investigators were changing cloze in the following ways within and between studies:

1) seven different scoring methods have been used
2) numerous deletion patterns have been tried
3) blank lengths have been modified
4) passage difficulties have been varied
5) test length has been changed
6) and a variety of different samples have been used.

These variables have been manipulated, consciously and unconsciously, in search of more effective ways to construct and interpret cloze tests. Generally speaking, variables one through five above have been purposefully manipulated or controlled in the second-language research. Variable six, the effect of different samples, has not been investigated sufficiently, which seems strange given that cloze procedure was originally shown to be very sample sensitive—so sensitive that readability grade levels could be established by using it (Taylor 1953).

In fact, sampling is an important consideration in many second language studies. After all, it is simple common sense that a sample of nonnative students taken at a university in Great Britain may be quite different from one taken at UCLA or in Papua New Guinea. Just such differences in samples exist in the studies cited above and this variable alone may have much to do with the wide variety of results. For example, Ebel (1979:290–91) has pointed out that the reliability of a set of test scores depends in part on the "range talent" in the group tested. In fact, restrictions in the range of talent can depress both reliability and validity coefficients in general (Shavelson 1981).

The purpose of this study, then, is to investigate the effects of differences in samples on cloze test results by addressing the following more specific research questions:

1) What are the effects of different ranges of talent on the apparent reliability and validity of cloze?
2) What is the strength of relationship between ranges of talent and the reliability and validity coefficients?
3) Do the results generalize to other cloze studies?

METHOD

Subjects

The samples in this study were all randomly selected (to be approximately equal in size) from larger university level populations and consisted of four groups which will be labeled as follows: 1) 1978 sample, 2) 1981 sample, 3) Winter 1982 sample and 4) Spring 1982 sample. The four samples (described in Table 1) differed in many ways, but it is particularly important to notice the way they differed in terms of estimated TOEFL score ranges (see last column). From these estimates, it is clear that the groups differed considerably in the ranges of talent represented in each.

Materials

The cloze passage under investigation here was adapted from *Man and His World* (Kurilecz 1969), an intermediate ESL reader. The passage was 399 words long and had an every 7th word deletion pattern for a total of fifty blanks. To provide context, two sentences were left intact (that is, without blanks) at the beginning of the passage and one at the end.

The measures used to calculate the criterion-related validity coefficients were all standardized (norm-referenced) English language placement or proficiency tests. They differed from sample to sample as follows: 1978 sample—UCLA English as a Second Language Placement Examination (ESLPE) (including listening and reading comprehension, dictation and structure subtests); 1981 sample—Guangzhou English Language Center (GELC) Placement Test (including listening and reading comprehension, as well as writing and structure subtests); Winter 1982 sample—*Test of English as a Second Language Practice Kit Number 1* (including listening, structure and written expression, as well as reading comprehension and vocabulary); Spring 1982 sample—UCLA ESLPE (a shorter version of the ESLPE used above without the dictation subtest).

Procedures

Exactly the same cloze passage was administered to each of the four samples and no more than two weeks separated its administration from that of the validity criterion measure. The cloze test was scored using two scoring methods: the *exact-answer method* (EX), wherein only the word found in the original passage is counted correct, and the *acceptable-word method* (AC), wherein any word acceptable to native speakers is counted correct. The latter method was based on the responses of 77 UCLA freshman composition students (Brown 1978).

Analyses

The descriptive test statistics in this study include the mean (x), standard deviation (S) and range. Cronbach alpha (r_{xx}) internal consistency reliabilities are also given along with criterion-related validity coefficients (r_{xy}). The latter were calculated by determining the correlation between the cloze tests and the criterion measure in question. All correlation coefficients reported in this study are Pearson product-moment coefficients.

Fisher *z* transformations were used whenever correlation coefficients were

Table 1: Sample Descriptions

Sample	Place	n	Sex		Academic Status			Nationalities	Major	Estimated TOEFL Range
			M	*F*	*Extension*	*Undergraduate*	*Graduate*			
1978	UCLA	55	45%	55%	18%	36%	46%	Numerous (See Brown 1978)	Numerous (See Brown 1978)	very wide (range = ±500)
1981	GELC	45	93%	7%	0	0	100%	Chinese	Engineering (100%)	259-578 (range = 319)
Winter 1982	GELC	45	78%	22%	0	0	100%	Chinese	Biochemistry (38%) Chemistry (33%) Biology (29%)	440-600 (range = 160)
Spring 1982	GELC	45	80%	20%	0	0	100%	Chinese	Engineering (44%) Agriculture (16%) Biology (7%) Other Sciences (33%)	435-515 (range = 80)

compared with standard deviations in order to correct for the non-symmetrical distribution of such coefficients. In general, this is necessary in order to draw correct inferences about sample correlation coefficients which are not near zero (Guilford and Fruchter 1973: 144-46).

RESULTS

Descriptive test characteristics are reported in Table 2 for the cloze test administered to the four different samples. These are the four samples described above. Remember that they were quite different in ranges of talent. These differences were also reflected on the cloze test in terms of test ranges (rows four and ten) and perhaps more accurately in the standard deviations (rows two and eight).

Table 2: Test Characteristics, Both Scoring Methods

Scoring Method	Statistic	1978 (n=55)	1979 (n=45)	Winter 1982 (n=45)	Spring 1982 (n=45)
EX	$\bar{x}$	15.00	23.33	24.36	21.78
	s	8.56	5.59	4.84	3.38
	low-high	0-33	9-31	14-36	13-27
	range	33	22	22	14
	r_{xx}	.90	.73	.68	.31
	r_{xy}	.88 (ESLPE)	.74 (GELC)	.59 (TOEFL)	.43 (ESLPE)
AC	$\bar{x}$	25.58	35.83	37.80	34.73
	s	12.45	6.71	4.48	4.07
	low-high	0-46	17-46	26-46	21-42
	range	46	29	20	21
	r_{xx}	.95	.83	.66	.53
	r_{xy}	.90 (ESLPE)	.79 (GELC)	.51 (TOEFL)	.40 (ESLPE)

The Effects of Different Ranges of Talent on the Reliability and Validity of Cloze.

In Table 3, the results are rearranged to illustrate that the reliability and validity coefficients decrease when the range of talent (as represented by standard deviation and test range) decreases. The deletion pattern, blank length, passage difficulty, test length and time allowed for the test were all held constant here while the sample range was systematically varied. The results indicate that a relationship exists between range of talent, and the various reliability and validity coefficients.

Another way of looking at this problem is to adjust the observed reliability coefficients for homogeneity of variances, or restrictions in range (after Magnusson 1967:75). When this is done, it turns out that the adjusted reliability coefficients are all between .95 and .96. Thus, the reliability coefficients would be virtually the same for all of the samples if it were not for the differences in variance. In short, the results here demonstrate that restrictions in range of talent do indeed depress the reliability and validity coefficients consistent with psychometric theory.

Table 3: Ranges of Talent in Relationship to Reliability and Validity of Cloze

Scoring Method and Sample	s	range	r_{xx}	r_{xy}
AC 1978	12.45	46	.95	.90
EX 1978	8.56	33	.90	.88
AC 1981	6.71	29	.83	.79
EX 1981	5.59	22	.73	.74
EX Winter 1982	4.84	22	.68	.59
AC Winter 1982	4.48	20	.66	.51
AC Spring 1982	4.07	21	.53	.40
EX Spring 1982	3.38	14	.31	.43

The Strength of Relationship between Ranges of Talent, and the Reliability and Validity Coefficients.

To evaluate the strength of association between range of talent, and reliability and validity coefficients, correlational analysis was performed. The correlation between standard deviations and reliability coefficients was found to be r = .97 for the two scoring methods combined. This indicates that about 93 percent (r^2) of the variation in reliability coefficients can be accounted for by knowing the standard deviations. Likewise, the strength of association between the standard deviations and validity coefficients was found to be r = .93 (r^2 = .86). In other words, the standard deviation seems to account for about 86 percent of the variation in validity coefficients.

In short, the results here indicate that variations in sample range, whether generated by the sample itself or the scoring method employed, strongly account for differences in the reliability and validity coefficients. This effect is so great that, depending on the sample and scoring method used, this cloze passage may appear to be one of the *best* passages ever reported (r_{xx} = .90; r_{xy} = .95 for AC 1978) or a hands-down loser of the *worst* (r_{xx} = .31; r_{xy} = .43 for EX Spring 1982).

Generalizability of the Results to Other Cloze Studies.

In answering this question, only those studies which provided clear and complete information (that is, standard deviation, reliability and validity coefficients) could be considered. In addition, only those based on 50-item passages scored by the EX and AC methods were included. The results of forty different sets of results are presented in Table 4. The correlation between the standard deviations and the reliability estimates throughout Table 4 was found to be .91. The squared value of this coefficient, .83, indicates that about 83 percent of the variation in reliability coefficients is explained by variation in the magnitude of the standard deviations. Likewise, the correlation between the standard deviations and the validity coefficients was .78 which shows that approximately 61 percent of the variation in validity coefficients is explained by variation in the standard deviations.

Notice that both of these relationships were found here even though five different deletion patterns and two scoring methods were combined.

In summary, the cloze literature to date indicates that cloze *may or may not* be highly reliable and valid as a norm-referenced test of overall second language proficiency. The results here indicate that this may be largely due to differences in the

Table 4: Reliability and Validity Related to Standard Deviation (Four Studies)

Study	Deletion	r_{xx}	S	Scoring Method	r_{xy}	(Criterion test)
Oller	7th word	.95	11.5	AC	.89	(ESLPE)
1972		.95	9.6	AC	.89	
		.92	8.8	EX	.87	
		.90	9.2	EX	.85	
		.90	6.6	AC	.80	
		.80	6.0	EX	.73	
Hinofotis	7th word	.85	7.3	AC	.79	(TOEFL)
1980		.61	2.1	EX	.71	
Alderson	12th word	.89	8.8	AC	.85	(ELBA)
1979		.88	8.0	AC	.78	
		.90	7.9	AC	.77	
		.80	6.5	EX	*.67	
		.79	6.3	EX	.70	
		.79	5.6	EX	.65	
	10th word	.87	8.1	AC	.83	(ELBA)
		.89	7.0	AC	.74	
		.80	6.4	AC	.74	
		.79	5.6	EX	.65	
		.69	4.8	EX	.57	
		.69	3.7	EX	*.79	
	8th word	.91	9.6	AC	.87	(ELBA)
		.87	7.9	AC	.77	
		.84	6.3	AC	.69	
		.79	6.3	EX	.70	
		.76	5.3	EX	.68	
		*.81	4.9	EX	.82	
	6th word	.88	8.2	AC	.88	(ELBA)
		.82	7.2	AC	*.67	
		*.84	5.8	AC	*.74	
		.80	5.8	EX	.86	
		.76	5.3	EX	.51	
		.53	4.3	EX	.53	
Present						
Study	7th word	.95	12.5	AC	.90	(ESLPE)
		.90	8.6	EX	.88	(ESLPE)
		.83	6.7	AC	.79	(GELC)
		.73	5.6	EX	.74	(GELC)
		.68	4.8	EX	.59	(TOEFL)
		.66	4.5	AC	.51	(TOEFL)
		.53	4.1	AC	.40	(ESLPE)
		.31	3.4	EX	.43	(ESLPE)

*Coefficient does not seem to fit the ordering.

way a given cloze passage relates to a given sample. This is consistent with psychometric theory and apparently is a factor in other cloze studies.

DISCUSSION

It should be emphasized that cloze is being viewed here as a norm-referenced test for purposes of placement or proficiency testing in ESL/EFL programs. Thus, the statistical concepts of reliability, validity, etc. are important considerations though they may seem a bit tedious to the hardworking teachers/ administrators in the field. To make these results more relevant to those very teachers, both theoretical *and* practical implications will be dicussed here.

Theoretical Implications

To the language testing specialist, the results here may seem obvious, based on knowledge of psychometric theory, to the point of being uninteresting. It may be, however, that the obvious has been overlooked in favor of the fashionable. Put in more scientific terms, the most parsimonious explanations of the phenomena we are observing in cloze testing may be found in the psychometric theory and statistical techniques being used. Or, the tools themselves may hold the clues to clear interpretations of the data.

Let us take for example a rather naive study (Brown 1980), the author of which will most definitely not sue for libel. In this study, four scoring methods were compared on the basis of reliability coefficients (ranging from .89 to .95), validity coefficients (ranging from .88 to .91) and other test characteristics. One conclusion drawn was that "the best *overall* scoring method is the AC method" (p. 316). While this conclusion seemed reasonable at the time based on previous research, information was available in that study, which should have been examined. For instance, the AC scoring method was nearly perfectly centered for the given sample ($\bar{x} = 25.58$ out of 50) and was the only scoring method for which the subjects were normally distributed (with the highest standard deviation of 12.45). The other three scoring methods produced distributions which were either negatively or positively skewed for the particular samples in question with correspondingly lower standard deviations. In addition, the same cloze passage administered to other samples in China has here been shown to have entirely different distributions in each of the samples with corresponding differences in the reliability and validity coefficients produced.

In short, the results obtained in Brown (1980) might have been quite different had intuition and good luck not guided the researcher to the particular passage and sample of subjects involved. Therefore, a more parsimonious and sensible hypothesis for differences in reliability and validity for different scoring methods (or deletion patterns, difficulty levels, etc.) might be that adjusting any and all variables which help to make a given cloze passage more appropriate for a given sample will correspondingly help to produce a test which is statistically more reliable and valid.

Furthermore, it appears that cloze is not necessarily a reliable, valid and easy to develop test of overall second language proficiency as is often believed (for example, Soudek and Soudek 1983). In fact, it is probably erroneous to say that cloze *is* anything; rather, it would be safer to take the position that cloze tests *are* a "family of item types" (Mullen 1979) which can tap the wide range in the universe of possible

language proficiency items (at least in the receptive/productive modes on written material).

It cannot be taken as a foregone conclusion that a given cloze test will be reliable and valid for a given sample because it would be a rare sample whose abilities spanned the entire range of possible items. Nevertheless, it is necessary to make decisions within samples that are more or less narrow in terms of ranges of talent. Therefore, it would seem that a cloze test should be made to *fit* a particular sample if decisions based on the results are to be responsible. This last necessity may preclude the notion that cloze tests are *easy to develop*.

Practical Implications

How can a cloze test be made to fit a given sample? First and foremost, cloze tests should be pretested like any other language tests so that the results can eventually provide clear interpretations. To this end, cloze items can be selected/fitted to a given sample in one of three ways: 1) the hit or miss method, 2) the modification method or 3) the well-tailored cloze method.

The hit or miss method. This shotgun approach to test development would involve selecting a relatively large number of tests, deleting every nth word and administering all of them to a sample of students representative of the group about which decisions would ultimately be made. After analyzing the results, that cloze test which seemed to produce the best distribution of scores could be selected for later decision making. In other words, the cloze passage which seemed to best center the sample (that is, produced a mean of about 50% correct) and which appeared most sensitive to the range of talent in that sample (that is, produced a high standard deviation) could be selected for later use with the entire group.

The modification method. To adopt this method, one cloze passage, which was thought to be intuitively *about the right level* for the group, could be developed and administered to a sample representative of the larger group. After analyzing the results using the EX scoring method, modifications could be made consistent with what has been found in the literature to date. For instance, if the cloze test in question was found to be much too difficult for the group (for example, produced a mean of 25% correct), it seems likely that lengthening the passage and increasing the distance between the blanks (from say every 7th word to every 11th word) would help to better center the scores. Alternatively, the mean could be somewhat artifically increased by using the AC scoring method. Using the AC method has also been shown to produce higher standard deviations in many but not all studies. The modified passage should then be readministered and reanalyzed to see that the desired effects had occurred and that the passage indeed fit the entire group.

The well-tailored cloze. It has been shown (Brown Unpublished ms.) that traditional test development techniques can be applied to a cloze test to increase the reliability of that instrument. Five different, but non-overlapping, every 7th word deletion pattern versions of one passage (50 items each) were administered to random samples of a group of Chinese students who had a very narrow range of talent. Analysis of the results produced item difficulty and discrimination indices for a pool of 250 possible items. From these items the *best* 50 were selected. In other words, those which had item difficulty levels most closely approximating .50 and the

highest discrimination indices were chosen. One restriction was placed on this selection process. The distance between items on the final version was to be no less than five words and no more than nine with an average of seven words. The new version of the test was then readministered to the same group after six weeks (to avoid *testing effect*) and found to be much more reliable than the original version with this same group. These results suggest that a cloze test can be tailored to fit a given group in much the same way that discrete-point tests have traditionally been developed (though perhaps without the same precision because of the differences in the context provided in the various versions involved).

Returning to the title of this study, and the overall question involved, it appears that a cloze is not a cloze is not a cloze. In fact, they appear to differ quite widely in effectiveness as norm-referenced instruments. This effectiveness in terms of reliability and validity, appears to be strongly related to how well a given cloze passage fits a given sample. Therefore, pretesting any cloze passage(s) seems absolutely essential so that an appropriate passage can be selected, modifications can be made or a passage can be tailored to fit a particular group of students. Taking some or all of these steps should help to produce a more highly reliable and valid norm-referenced instrument. Only then can adequately responsible decisions be based on the scores of our students on such a test.

REFERENCES

Alderson, J. Charles. 1979. Scoring procedures for use on cloze tests. In *On TESOL 79*, Carlos A. Yorio, Kyle Perkins and Jacqueline Schachter (Eds.). Washington, D.C.: TESOL.

Alderson, J. Charles. 1980. Native and non-native speaker performance on cloze tests. *Language Learning* 30(1):59-76.

Brown, James D. 1978. Correlational study of four methods for scoring cloze tests. MA thesis, University of California Los Angeles.

Brown, James D. 1980. Relative merits of four methods for scoring cloze tests. *Modern Language Journal* 64(3):311-317.

Brown, James D. 1983. A closer look at cloze: part II—reliability. In *Issues in language testing research*, John W. Oller, Jr. (Ed.). Rowley, Massachusetts: Newbury House Publishers, Inc.

Brown, James D. Undated. Well-tailored cloze. Manuscript.

Conrad, Christine A. 1970. The cloze procedure as a measure of English proficiency. MA thesis, University of California Los Angeles.

Darnell, Donald K. 1970. Clozentropy: a procedure for testing English language proficiency of foreign students. *Speech Monographs* 37:36-46.

Ebel, Robert L. 1979. *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Guilford, Joy Paul and Benjamin Fruchter. 1973. *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill Book Company.

Hinofotis, Frances Butler. 1980. Cloze as an alternative method of ESL placement and proficiency testing. In *Research in language testing*, John W. Oller, Jr. and Kyle Perkins (Eds.). Rowley, Massachusetts: Newbury House Publishers, Inc.

Irvine, Patricia, Parvin Atai and John W. Oller, Jr. 1974. Cloze, dictation, and the test of English as a foreign language. *Language Learning* 24(2):245-252.

Jonz, Jon. 1976. Improving on the basic egg: the M-C cloze. *Language Learning* 26 (2):255-265.

Kurilecz, Margaret. 1969. *Man and his world: a structured reader*. New York: Crowell.
Magnusson, David. 1967. *Test theory*. Reading, Massachusetts: Addison-Wesley Publishing Company.
Mullen, Karen. 1979. More on cloze tests as tests of proficiency in English as a second language. In *Concepts in language testing: some recent studies*, Eugene J. Briere and Frances Butler Hinofotis (Eds.). Washington, D.C.: TESOL.
Oller, John W., Jr. 1972a. Dictation as a test of ESL proficiency. In *Teaching English as a second language: a book of readings*, Harold B. Allen and Russell N. Campbell (Eds.). New York: McGraw-Hill Book Company.
Oller, John W., Jr. 1972b. Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal* 56(3):151–158.
Oller, John W., Jr. 1975. Research with cloze procedure in measuring the proficiency of non-native speakers of English: an annotated bibliography. ERIC/CLL, ED 104 154.
Oller, John W., Jr. 1979. *Language tests at school: a pragmatic approach*. London: Longman Group Ltd.
Oller, John W., Jr. and Nevin Inal. 1971. A cloze test of English prepositions. *TESOL Quarterly* 5(4):315–326.
Pike, Lewis W. 1973. *An evaluation of present and alternative item formats for use in the test of English as foreign language*. Princeton, New Jersey: Educational Testing Service.
Shavelson, Richard J. 1981. *Statistical reasoning for the behavioral sciences*. Boston, Massachusetts: Allyn and Bacon, Inc.
Soudek, Miluse and Lev I. Soudek. 1983. Cloze after thirty years: many new uses in language teaching—a checklist of sources. Paper presented at the 17th Annual TESOL Convention, Toronto, Canada, March 15–20, 1983.
Taylor, Wilson L. 1953. Cloze procedure: a new tool for measuring readability. *Journalism Quarterly* 30:414–438.